NBER WORKING PAPER SERIES

KNOWLEDGE, TESTS, AND FADEOUT IN EDUCATIONAL INTERVENTIONS

Elizabeth U. Cascio
Douglas O. Staiger

Working Paper 18038
http://www.nber.org/papers/w18038

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
May 2012

We thank Kevin Lang and Thomas Lemieux for the conversations that motivated this research, and Raj Chetty, Caroline Hoxby, Kevin Lang, Larry Katz, Jesse Rothstein, and seminar participants at the NBER Education Program meetings for helpful comments. We are also grateful for the outstanding research assistance of Sarah Cohodes. This work was supported by the Institute of Education Sciences, U.S. Department of Education grant [R305C090023] to the President and Fellows of Harvard College. Cascio also gratefully acknowledges funding from the National Academy of Education and the National Academy of Education/Spencer Postdoctoral Fellowship Program and a Rockefeller-Haney Fellowship from Dartmouth College. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Knowledge, Tests, and Fadeout in Educational Interventions
Elizabeth U. Cascio and Douglas O. Staiger
NBER Working Paper No. 18038
May 2012
JEL No. I20,I21,I28,J24

ABSTRACT

Educational interventions are often evaluated and compared on the basis of their impacts on test scores. Decades of research have produced two empirical regularities: interventions in later grades tend to have smaller effects than the same interventions in earlier grades, and the test score impacts of early educational interventions almost universally "fade out" over time. This paper explores whether these empirical regularities are an artifact of the common practice of rescaling test scores in terms of a student's position in a widening distribution of knowledge. If a standard deviation in test scores in later grades translates into a larger difference in knowledge, an intervention's effect on normalized test scores may fall even as its effect on knowledge does not. We evaluate this hypothesis by fitting a model of education production to correlations in test scores across grades and with college-going using both administrative and survey data. Our results imply that the variance in knowledge does indeed rise as children progress through school, but not enough for test score normalization to fully explain these empirical regularities.

Elizabeth U. Cascio
Department of Economics
Dartmouth College
6106 Rockefeller Hall
Hanover, NH 03755
and NBER
elizabeth.u.cascio@dartmouth.edu

Douglas O. Staiger
Dartmouth College
Department of Economics
HB6106, 301 Rockefeller Hall
Hanover, NH 03755-3514
and NBER
douglas.staiger@dartmouth.edu

I. Introduction

Educational interventions are often evaluated and compared on the basis of their impacts on test scores. Decades of research in this vein have produced two seemingly contradictory empirical regularities. An intervention in a later grade frequently has a smaller impact on test scores than the same intervention in an earlier grade.[1] At the same time, the test score impacts of early educational interventions almost universally "fade out" as participants progress through school.[2] On one hand, the scope for education to increase knowledge would appear to be greater when children are young; on the other, the effects of early intervention on knowledge do not appear to last.

In this paper, we investigate the extent to which these regularities are an artifact of the common practice of rescaling test scores in terms of a student's position in a widening distribution of knowledge. If the variance in knowledge rises as children progress through school, a standard deviation in test scores in a higher grade will correspond to a greater difference in knowledge than a standard deviation in test scores in a lower grade.[3] Conversely, an educational intervention that has the same impact on knowledge will appear to have a smaller impact on normalized test scores when the intervention occurs in a later grade. Likewise, an early intervention's effect on normalized test scores may decline in magnitude over time even if its impact on knowledge persists in full.

Though this hypothesis is intuitive, evaluating it is difficult, since knowledge has no natural scale. Indeed, the near universal practice of re-expressing achievement in distributional terms is a practical response to this fact, as tests (and thereby test scales) often change as children in a target population age. Given this, our empirical approach works off of several patterns that we might

[1] Such interventions include test-based accountability (e.g., Dee and Jacob, 2011), charter schools (Dobbie and Fryer, 2011), and teachers (Kane, Rockoff, and Staiger, 2008), among others.

[2] Such early interventions include compensatory preschool education (see Almond and Currie (2011) for a review), class size reduction (Krueger and Whitmore, 2001, Chetty, et al., 2011), and teacher assignment (Kane and Staiger, 2008; Jacob, Lefgren, and Sims, 2008; Rothstein, 2010, Chetty, Friedman, and Rockoff, 2011).

[3] The same intuition holds if we were to consider rescaling test scores in percentiles rather than in standard deviation units. We focus on standard deviation units here because the linearity of the underlying transformation considerably simplifies the analysis.

expect to see in data if the variance of knowledge were increasing across grades. In particular, we expect test scores in consecutive grades to be more correlated with one another at higher grades than at lower grades, since at higher grades, dramatic shifts in the accumulated knowledge distribution should be less likely. Similarly, test scores in later grades should be more strongly correlated with long-term, non-test outcomes, like educational attainment, than test scores in earlier grades.[4]

More formally, we lay out a simple model of education production, motivated by econometric specifications frequently employed in the empirical literature, in which test scores are noisy measures of knowledge and the stock of knowledge accumulates over time as new knowledge is added each year and past knowledge potentially depreciates. The variance of accumulated knowledge in each grade is a function of model parameters. The correlations in test scores across grades and with long-term outcomes are as well, allowing us to estimate the model parameters using minimum distance methods. We use these parameter estimates to predict how the variance in knowledge evolves across grades, and in turn, the extent to which rising variance in the distribution of knowledge contributes to relatively weak contemporaneous impacts of educational interventions in later grades and to fadeout in early interventions. We fit the model using student-level administrative data on test scores and college-going from Charlotte-Mecklenburg Schools and survey data from the 1979 National Longitudinal Survey of Youth. Neither data source is ideal for our analysis, but they are complementary and offer independent estimates.

Our estimates are similar across data sets, and imply that the variance in accumulated knowledge increases by between 37 and 56 percent from the end of kindergarten to the end of high school. Due to this widening distribution of knowledge, we find that a modest amount of the apparent decline in the impact of interventions in later grades is an artifact of normalizing test

[4] Increasing test reliability across grades could create similar patterns. Our empirical approach allows us to estimate test reliability in a flexible way to account for this possibility.

scores. For example, the impact of attending a small class in kindergarten on normalized test scores would be 12 to 17 percent lower by eighth grade even if the effect of a small kindergarten class on knowledge persisted in full. Similarly, the impact of an effective teacher on students' normalized test scores would be 6 to 10 percent less if she were assigned to teach eighth grade rather than third grade. While the statistical artifact by itself cannot explain why over half of the impact of early educational interventions or having an effective teacher fades in a few years, it is still substantial, particularly from the perspective of policymakers seeking to compare interventions across grades.

Much of the evaluation research cited above does not directly acknowledge the comparability of tests across grades. In the economics literature, the issue of test scaling has received most attention in the context of understanding teacher impacts on student achievement (Cawley et al., 1999; Ballou, 2009; Lang, 2010; Barlevy and Neal, 2011). Lang (2010), in particular, proposes that the apparent fadeout of teacher effects in later grades may be a statistical artifact of using normalized test scores. Our paper formalizes and empirically tests Lang's idea. In the education literature, the issue of test scaling has received most attention in the context of understanding achievement gaps between groups (Selzer et al., 1994; Reardon, 2007; Reardon and Robinson, 2007; Bond and Lang, 2011). Our results suggest that estimates based on standardized tests will understate racial gaps in knowledge for later grades, and therefore understate growth in racial gaps as children age (see, for example, Fryer (2011)). Finally, in the psychometric literature, the issue of test scaling has been addressed using Item Response Theory (IRT) (Hambleton, Swaminathan and Rogers, 1991). The IRT model has been used to align test scores across grades and over time, but often yields implausible scales in practice (Ballou, 2009) and is not anchored to adult outcomes (Cunha and Heckman, 2008). We provide an alternative approach to IRT, based on commonly accepted scales with well documented relationships to adult outcomes (Chetty et al. 2011; Chetty, Friedman, and Rockoff, 2011).

II. Education Production

Our objective is to estimate how the variance in knowledge changes as children progress through school. The challenge for our analysis is that knowledge is unobservable and has no natural scale. To overcome this obstacle, we model the data-generating process for correlations between test scores across grades and with a longer-term, non-test outcome, which can be observed. The parameters of this model imply how the variance of knowledge evolves across grades and can be estimated using minimum distance methods, as we discuss in Section III.

A. Model

We begin by modeling the accumulation of knowledge. We assume that each child starts school with some knowledge stock, and that education augments knowledge in ways that could be transitory or persistent (as in Jacob, Lefgren, and Sims, 2008). For example, rote memorization or "teaching to the test" might increase knowledge temporarily,[5] while deeper learning will have more persistent effects.

More formally, suppose that the knowledge of child *i* in grade *t*, g_{it}, can be decomposed into long-term (*L*) and short-term (*S*) components:[6]

$$g_{it} = g_{it}^{L} + g_{it}^{S}, \tag{1}$$

We model the short-term component, g_{it}^{S}, as a shock to knowledge in *t* that does not carry over to later grades, ω_{it}:

$$g_{it}^{S} = \omega_{it}, \tag{2}$$

[5] There are other reasons that an intervention might have appear to have transitory effects on knowledge. For example, if an intervention primarily affects understanding of material that is not tested in subsequent years, it would manifest as a transitory increase in test performance, even though this understanding might persist in full. Alternatively, interventions may have true effects on knowledge that fade rapidly. For example, mixing of treatment and control students in the same classrooms after a classroom-level intervention has ended may promote rapid reversion toward the mean in student performance on material that students learn in school, while student-level test performance may fade less slowly as it also includes more persistent student-level factors.

[6] For exposition, we consider one cohort of students which progresses through school one grade per year, so period is synonymous with grade. We discuss our approach to handling grade repetition in the data section below.

and the long-term component, g_{it}^{L}, as comprised of an individual's endowment, μ_i, and knowledge accumulated since starting school, α_{it}:

(3) $$g_{it}^{L} = \mu_i + \alpha_{it}.$$

We model the latter as an AR(1) process:

(4) $$\alpha_{it} = \delta\alpha_{it-1} + v_{it},$$

where v_{it} represents the innovation to long-term knowledge in grade *t*, and the parameter δ captures the degree of persistence of innovations from previous grades. Beyond constraining δ to be positive, we place no restrictions on its value in our estimation: innovations to long-term knowledge can decay ($0 < \delta < 1$) or grow ($\delta > 1$) as a child ages. In the special case where $\delta = 1$ (a random walk), α_{it} is simply the sum of all innovations v_{it} from school entry through grade *t*.

We assume that the ω_{it} and the v_{it} are drawn independently over time, independently of μ_i, and independently of one other. These assumptions have several implications for interpretation of model parameters.[7] First, equation (4) should be thought of as a linear projection, where the innovation to long-term knowledge is uncorrelated with accumulated knowledge by definition. Thus, any factors that may make innovations to long-term knowledge positively correlated over time for a given individual, like ability tracking, are subsumed in δ. That is, if students are tracked on the basis of early innovations, knowledge differences will appear more persistent, or the value of δ will rise.

Second, the assumption that ω_{it} and v_{it} are independent of μ_i affects interpretation of the endowment variance, σ_{μ}^{2}. To see this, it is useful to rewrite equation (1) as:

[7] Note that equations (1) through (4) imply that g_{it} is composed of a student fixed effect plus an AR(1) plus white noise. It is well known (Granger and Morris, 1976) that any AR(1) + white noise process can be equivalently represented by an ARMA(1,1). Thus, we could equivalently capture the persistent and transitory shocks by modeling g_{it} as the sum of a student fixed effect and an ARMA(1,1), where the MA term captures the transitory component of each shock.

(1')
$$g_{it} = \mu_i + \delta\alpha_{it-1} + v_{it} + \omega_{it} ,$$

substituting for g_{it}^{S} and g_{it}^{L} based on equations (2) through (4). For the sake of argument, suppose that μ_i represents *i*'s knowledge at the start of school. If children with a higher stock of knowledge at school entry have a firmer foundation for learning – and thus learn more in the early grades – it would then be the case that μ_i is positively correlated with v_{it} (and ω_{it}). By assuming that learning is orthogonal to the endowment, we thus subsume in μ_i some of the systematic differences across children in the rate of learning once they enter school. If education accentuates pre-existing differences in knowledge across children, as hypothesized above, σ_{μ}^{2} will exceed the variance in knowledge at the start of school.

The variances of the innovations may change across grades. In practice, we cannot easily statistically distinguish ω_{it} from testing noise, so we model the variance of ω_{it} jointly with the variance of testing error (see below). As for v_{it}, we begin with a specification where $\text{var}(v_{it})$ is fixed, then allow the variance of v_{it} to change smoothly across grades according to:

(5)
$$\text{var}(v_{it}) = \beta_v \,\text{var}(v_{it-1}),$$

where $\beta_v > 0$, and $\beta_v > 1$ $(\beta_v < 1)$ implies higher (lower) innovation variance in later grades.

Test scores, y_{it}, are assumed to be noisy measures of knowledge at a given point in time:

(6)
$$y_{it} = g_{it} + \varepsilon_{it} ,$$

where ε_{it} is a mean zero measurement error drawn independently of g_{it} and independently over time. ε_{it} represents idiosyncratic features of the test environment that contribute to an individual scoring unusually well or unusually poorly (e.g., a "barking dog"). It is useful to parameterize the error variance, $\text{var}(\varepsilon_{it})$, alongside the variance in short-term knowledge, $\text{var}(\omega_{it})$, using a reliability ratio, λ_{y_t} :

$$\lambda_{y_t} = \frac{\operatorname{var}\left(g_{it}^{L}\right)}{\operatorname{var}\left(g_{it}^{L}\right) + \operatorname{var}\left(\omega_{it}\right) + \operatorname{var}\left(\varepsilon_{it}\right)}. \tag{7}$$

λ_{y_t} is distinct from what is usually meant by reliability of a test, e.g., test-retest reliability. Instead, λ_{y_t} is the share of total variance in test scores accounted for by the variance in long-term knowledge. Thus, while λ_{y_t} will be lower for tests with lower test-retest reliability, it will also be lower when a higher share of knowledge is short-term. Because this share might change, we allow λ_{y_t} to vary across grades in our estimation.[8] Reported test reliabilities, where available, should provide an upper bound on these parameters.

Finally, we assume that some longer-term non-test outcome, w_i – in our case, college going – is a noisy measure of the long-term knowledge at the end of secondary school, period *T,* g_{iT}^{L}. Without loss of generality, we normalize w_i so that the coefficient on g_{iT}^{L} is equal to one:

$$w_i = g_{iT}^{L} + u_i. \tag{8}$$

u_i is a mean zero independent error term that represents all factors orthogonal to g_{iT}^{L} that contribute to variation in *w*. Again, it is convenient to parameterize the error variance with a reliability ratio:

$$\lambda_w = \frac{\operatorname{var}\left(g_{iT}^{L}\right)}{\operatorname{var}\left(g_{iT}^{L}\right) + \operatorname{var}\left(u_i\right)} \tag{9}$$

We expect estimates of λ_w to be lower than estimates of λ_{y_t}. Long-term outcomes like college-going – while a function of the knowledge that tests are designed to measure – almost certainly depend on much more ("non-cognitive skills," for example).

[8] In our data, we observe multiple cohorts in the same grade. We therefore also allow λ_{y_t} to vary over time to account for changes in test-retest reliability stemming from changes in the test.

The model of education production represented in equations (1) through (9) is consistent with standard econometric specifications commonly employed in the empirical literature. For example, our model implies the following value-added specification for normalized test scores, $\tilde{y}_{it}$:

$$\tilde{y}_{it} = \left(\frac{\delta a_{t-1} - a_t}{b_t} \right) + \left(\frac{1-\delta}{b_t} \right) \mu_i + \delta \left(\frac{b_{t-1}}{b_t} \right) \tilde{y}_{it-1} + \left(\frac{1}{b_t} \right) v_{it} + \left(\frac{1}{b_t} \right) \left((\omega_{it} + \varepsilon_{it}) - \delta(\omega_{it-1} + \varepsilon_{it-1}) \right)$$

where a_t and b_t represent the mean and standard deviation in period t test scores, respectively. Thus, our model implies that period t normalized test scores are a linear function of fixed student characteristics (captured in μ_i), last period's test score (the coefficient on which captures both true knowledge decay and changes in the variance of knowledge), the current period innovation (which might include teacher and peer effects), and a serially correlated measurement error. The coefficients on these variables are time-varying if the standard deviation of test scores changes over time. This specification is similar to standard teacher value-added specifications, such as those that have recently been validated using both experimental and observational data (Kane and Staiger, 2008; Chetty, Friedman, and Rockoff, 2011). In addition, we follow Jacob, Lefgren, and Sims (2008) in the way our model specifies short-term shocks to knowledge (ω_{it}). This specification allows for rapid fadeout of interventions (e.g., of teacher effects) as found by the previously cited papers and others (Rothstein, 2008; McCafferty et al., 2004; Lockwood et al., 2007).

B. Implications

What are the implications of the model for measurement and interpretation of the impacts of educational intervention? Let a randomly-assigned intervention in period t on test performance in period $t+k$ ($k \geq 0$) be represented by θ_{t+k}^{t} standard deviations. That is,

$$\theta_{t+k}^{t} \equiv E\left[\tilde{y}_{it+k,t_1}\right] - E\left[\tilde{y}_{it+k,t_0}\right],$$

where the subscripts t_1 and t_0 represent the treatment and control groups in the period t intervention, respectively, and $\tilde{y}_{it+k}$ represents test scores in standard deviation units, i.e.,

$$\tilde{y}_{it+k} = \frac{y_{it+k} - E(y_{it+k})}{\sqrt{\text{var}(y_{it+k})}}.$$

As described above, a number of early interventions have found that $\theta^{t}_{t+k} < \theta^{t}_{t}$, or that effects on test scores fade out over time. Fadeout is widely interpreted as showing that the effects of educational intervention on knowledge rapidly decay. Our model implies that fadeout can also be a statistical artifact that arises from the practice of re-expressing student test scores in distributional terms. The same statistical artifact can generate the finding that $\theta^{t+k}_{t+k} < \theta^{t}_{t}$ – that educational interventions have larger contemporaneous effects on test scores when they take place in earlier grades – even when they have identical effects on knowledge.

To see the first result, note that under the statistical properties of the model and the assumption of random assignment, the contemporaneous effect of a period t intervention, θ^{t}_{t}, and its effect k grades later, θ^{t}_{t+k}, are related by the following expression:

(10) $$\frac{\theta^{t}_{t+k}}{\theta^{t}_{t}} = \delta^{k} \cdot (1-\tau) \cdot \sqrt{\frac{(1/\lambda_{y_t})\text{var}(g^{L}_{it})}{(1/\lambda_{y_{t+k}})\text{var}(g^{L}_{it+k})}},$$

where τ represents the share of the intervention's contemporaneous effect that does not carry over to later grades:

$$\tau = \frac{E(\omega_{it,1}) - E(\omega_{it,0})}{[E(v_{it,1}) - E(v_{it,0})] + [E(\omega_{it,1}) - E(\omega_{it,0})]}.$$

The last term in (10) represents the change between t and $t+k$ in the variance in test scores, which depends upon changes in the variance of long-term knowledge and changes in test reliability. This last part arises from the practice of normalizing test scores, and so is a statistical artifact. The common interpretation of fadeout implicitly assumes that this term is equal to one for all *t* and *k*. Under this assumption, fadeout only arises when an intervention's effect on knowledge decays, and fadeout is more rapid the higher is τ and the lower is δ. Equation (10) implies, however, that there are scenarios (as when $\delta \geq 1$ and $\tau = 0$) where fadeout could be observed without any knowledge decay. The statistical artifact may also be present when an intervention's effect on knowledge does indeed decay ($\delta < 1$ and/or $\tau > 0$).

This same statistical artifact will also drive a wedge between the contemporaneous test score impacts of two interventions that have the same impacts on knowledge in different grades. Let $\widetilde{\theta}_t^{\,t}$ and $\widetilde{\theta}_{t+k}^{\,t+k}$ represent the contemporaneous test score impacts of interventions in *t* and *t* + *k* that have identical impacts on knowledge.[9] Then $\widetilde{\theta}_t^{\,t}$ and $\widetilde{\theta}_{t+k}^{\,t+k}$ are related by the following expression:

$$(11) \qquad \frac{\widetilde{\theta}_{t+k}^{\,t+k}}{\widetilde{\theta}_t^{\,t}} = \sqrt{\frac{\left(1/\lambda_{y_t}\right)\mathrm{var}\left(g_{it}^{L}\right)}{\left(1/\lambda_{y_{t+k}}\right)\mathrm{var}\left(g_{it+k}^{L}\right)}}\,.$$

If the variance in knowledge is increasing more quickly than test reliability, $\widetilde{\theta}_{t+k}^{\,t+k} < \widetilde{\theta}_t^{\,t}$: the contemporaneous impact of the period *t* +*k* intervention on test scores will be less than that of the intervention in period *t*, despite the interventions' identical impacts on knowledge.

The goal of our empirical analysis is to estimate the statistical artifact. To do so, we must make several assumptions. The first assumption concerns test reliability. Equations (10) and (11) show that the statistical artifact is a function not only of the variance in accumulated knowledge, but also of test reliability, which varies across grades, over time, and across datasets, as shown below. Because changes in test reliability across grades for any given cohort will also be unique to any given intervention, we take as our starting point the assumption of unchanging test reliability ($\lambda_{y_t} = \lambda_{y_{t+k}}$ for all *t* and *k*). If test reliability is in fact (weakly) increasing across grades – which generally appears

[9] That is, $E\lfloor g_{it,t_1}\rfloor - E\lfloor g_{it,t_0}\rfloor = E\lfloor g_{it+k,t+k_1}\rfloor - E\lfloor g_{it+k,t+k_0}\rfloor$.

to be the case – this assumption yields an upper bound on the magnitude of the statistical artifact. We show this below by re-estimating the statistical artifact using our estimates of test reliability for the grades represented in our data.

Under the assumption of constant test reliability, the statistical artifact reduces to $\sqrt{\operatorname{var}\left(g_{it}^{L}\right)/\operatorname{var}\left(g_{it+k}^{L}\right)}$. As noted, estimation of this quantity would generally be impossible since the variance of knowledge cannot be observed. However, the model yields an expression for it:

(12) $$\operatorname{var}\left(g_{it}^{L}\right)=\sigma_{\mu}^{2}+\operatorname{var}\left(\alpha_{it}\right),$$

where the variance in cumulative "value-added" is:

(13) $$\operatorname{var}\left(\alpha_{it}\right)=\sigma_{v_1}^{2}\beta_{v}^{t-1}\sum_{j=0}^{t-1}\delta^{2j}\beta_{v}^{-j},$$

where $\sigma_{v_1}^{2}$ represents the variance of the first period innovation to knowledge, and recall that β_v characterizes how $\operatorname{var}\left(v_{it}\right)$changes over time (equation (5)). $\operatorname{var}\left(\alpha_{it}\right)$will be larger and grow faster across grades the larger is δ – or the less knowledge decays over time – and the larger is β_v – or with persistence or expansion of the innovation variance.

Together, equations (12) and (13) show that only ratios in the variance of long-term knowledge can be interpreted. For example, we could set the value of either σ_{μ}^{2} or $\sigma_{v_1}^{2}$, and the values undertaken by $\sqrt{\operatorname{var}\left(g_{it}^{L}\right)/\operatorname{var}\left(g_{it+k}^{L}\right)}$ would be unchanged. In practice, we set $\sigma_{v_1}^{2}=1$, and σ_{μ}^{2} is measured in multiples thereof. The next section discusses our approach to estimating these model parameters.

III. Estimation and Identification

Estimation of the statistical artifact requires estimation of model parameters. Our estimation strategy takes advantage of the fact that the model characterizes the data generating process for correlations between test scores across grades and between test scores and the non-test outcome. We choose model parameters to minimize the distance between model predictions and actual correlations observed in the data.

Under the model, the correlation between test scores in periods *t* and *t*+*k* is given by:[10]

$$\rho_{y_t y_{t+k}} = \sqrt{\lambda_{y_t}\lambda_{y_{t+k}}} \cdot \frac{\sigma_\mu^2 + \delta^k \operatorname{var}(\alpha_{it})}{\sqrt{\left(\sigma_\mu^2 + \operatorname{var}(\alpha_{it})\right)\left(\sigma_\mu^2 + \operatorname{var}(\alpha_{it+k})\right)}}, \tag{14}$$

where $\operatorname{var}(\alpha_{it})$ is given in equation (13) and all other parameters have been previously defined. Analogously, the correlation between test scores in *t* and college-going is:

$$\rho_{y_t w} = \sqrt{\lambda_{y_t}\lambda_w} \cdot \frac{\sigma_\mu^2 + \delta^{T-t} \operatorname{var}(\alpha_{it})}{\sqrt{\left(\sigma_\mu^2 + \operatorname{var}(\alpha_{it})\right)\left(\sigma_\mu^2 + \operatorname{var}(\alpha_{iT})\right)}} \tag{15}$$

Comparison of equation (15) to equation (14) shows that the correlation between period *t* test scores and the long-run outcome will be less than the correlation between test scores in period *t* and at the end of secondary school, period *T*, if the long-term outcome is a lower reliability measure of period *T* knowledge.

It is useful to consider several special cases for intuition as to how the model parameters are identified. When $\delta = 0$, so that there is no long-term component to knowledge ($\operatorname{var}(\alpha_{it}) = 0$) and value-added through education decays immediately, $\rho_{y_t y_{t+k}} = \sqrt{\lambda_{y_t}\lambda_{y_{t+k}}}$ and $\rho_{y_t w} = \sqrt{\lambda_y \lambda_w}$: the correlations depend only on reliability ratios, and knowledge itself is determined only by child endowments. Under the assumption that all tests are equally reliable, moreover, the correlations between test scores are exactly the same for all *t* and *k*, and the correlations between test scores and long-term outcomes do not vary with *t*.

[10] In some specifications, we allow test reliability to vary across both grades and years, not just across grades as shown here. We also allow the endowment variance to vary across cohorts. Rationales are given in Section IV (Data).

Continuing with the assumption of constant reliability, consider the case where $\delta = 1$ and $\beta_v = 1$, or where knowledge accumulation is a random walk and the innovation variance is constant. Because the variance in knowledge is rising across grades, both $\rho_{y_t y_{t+k}}$ and $\rho_{y_t w}$ fall below the values that would be predicted by reliability alone. The correlations also have gradients in *t*, the grade in which the (first) test is administered. In particular, for any given *k*, $\rho_{y_t y_{t+k}}$ is rising in *t*: intuitively, test scores are more correlated in higher grades, when more knowledge has been accumulated. $\rho_{y_t y_{t+k}}$ is also falling in *k*: the correlations between test scores are weaker the further apart the tests are taken. Similarly, $\rho_{y_t w}$ is greater the later in the school career the test is administered.

To clarify, Figures I and II plot simulated correlations from these special cases. Figure I plots the $\rho_{y_t y_{t+k}}$. The x-axis gives the grade that the first test was administered (*t*). The number of grades ahead the second test is administered (*k*) is represented by the number adjacent to each line; the correlations themselves are represented with solid dots. Thus, the correlation between third and fourth grade test scores is the point on the line labeled with a one (*k*=1) corresponding to grade 3 (*t*=3) on the x-axis. Figure II plots the $\rho_{y_t w}$ and is arranged in a similar fashion, with the grade of the test on the x-axis. We set the reliability ratios and the variance of the endowment to be roughly comparable to what we estimate in the CMS data.[11] The figures illustrate the predictions described above: the correlations are unchanging when $\delta = 0$ (Panel A), and increasing in *t* and falling in *k* when $\delta = 1$ and $\beta_v = 1$ (Panel B).

Figures like these also provide a useful vehicle for understanding the implications of further variation in these model parameters. For example, consider a case where $\delta = 1$ and $\beta_v = 0.9$, as shown in Panel C. Here, knowledge continues to be a random walk, but the innovations to

[11] As in our estimation, we also assume that *t*=1 represents kindergarten and *T*=13. We are also setting $\sigma^2_{v_1} = 1$ in these simulations, as we do in our estimation, with σ^2_{μ} rescaled accordingly.

knowledge are higher variance earlier in a child's school career. In this case, the correlations between test scores strengthen, particularly in higher grades: that is, the $\rho_{y_t y_{t+k}}$ are much more similar regardless of *k*, particularly at high values of *t* (Figure I). For example, the correlations between test scores in grades 3 and 4, grades 3 and 5, grades 3 and 6, etc. are all stronger and more similar to each other than was the case in Panel B. Similarly, there is less of a gradient in *t* in the correlation between test scores and the long-term outcome (Figure II). These predictions are intuitive, since more knowledge is accumulated relatively early. When $\beta_v > 1$, or when knowledge is accumulated relatively late, the opposite happens: the correlations weaken, and the gradients of the $\rho_{y_t y_{t+k}}$ in *k* and the $\rho_{y_t w}$ in *t* steepen.

Finally, Panel D of each figure shows what happens when $\delta = 0.9$ and $\beta_v = 1$. Here, the variance in the long-term innovation remains the same throughout the school career, as was the case in Panel B, but not all of this innovation persists. The primary consequence of knowledge decay is to reduce the gradient of $\rho_{y_t y_{t+k}}$ in *t* for any *k*. For example, the correlations between test scores in grades 3 and 4, between grades 4 and 5, between grades 5 and 6, etc. are more similar now than they were in the baseline case. Related, the correlations between test scores and long-term outcome do not rise as quickly in *t*. Again, this is intuitive, since differences in knowledge accumulated between any two grades will be less than if all knowledge were to persist. Further, variation in δ appears to add more curvature to these relationships: for example, the correlation between test scores and long-run outcomes – either *w* or much later test scores – can fall across early grades, if knowledge fades more rapidly than it is accumulating.

Thus, both β_v and δ affect the speed through which knowledge is accumulated – albeit through different mechanisms – and these differences manifest in different ways in the data. For example, variation in β_v is manifested more in the drop off in the test score correlations in *k*, while variation in δ is manifested more in how the $\rho_{y_t y_{t+k}}$ change with *t* for any given *k*. Data on *w* are thus not needed to estimate the model, but long-term outcomes are quite useful for identification purposes, since β_v and δ are in practice identified through fairly subtle differences in the patterns of these correlations. The correlation of test scores from any grade with long term outcomes is proportional to the correlation with knowledge in the final grade, and therefore is similar to observing test scores in 12th grade for all students (e.g., the lower envelope of the lines plotted in Figure I). In practice, identification of these parameters is complicated further by changing test reliability across grades, which adds noise to the relationships plotted in Figures I and II.

IV. Data

Given the discussion above, the ideal data set for estimating the model offers frequent observations over a wide span of grades on a large number of individuals, allowing each correlation to be estimated precisely. Neither of our data sources meet both of these criteria, but the two are complementary and provide independent estimates. This section describes samples and key variables from each and discusses their relative strengths and limitations. In both samples, we focus on math test scores.

A. Data from Charlotte-Mecklenburg Schools

Our administrative data are from Charlotte-Mecklenburg Schools (CMS), a large school district in North Carolina. CMS is one of the few school districts in the country with data sufficient for our analysis: similar tests (the North Carolina end of grade (EOG) exams in math) were administered in many consecutive years (1999 to 2009) and consecutive grades (grades 3 to 8), and far enough in the past that we can observe correlations between standardized test scores with a longer-term outcome – whether an individual enrolled in a four-year college in his first year out of high school, collected through the National Student Clearinghouse (NSC) – for several cohorts of students (those in third grade between spring of 1994 and spring of 1998). Another benefit of the

CMS data is that the size of the district makes the correlations quite precise, which helps to reduce noise in our estimates. The correlations are estimated on average using about 5400 students per cohort (grade-year).

One drawback of the CMS data, however, is that for no one cohort can we observe all possible correlations. Appendix Table I provides a description of the available data by cohort, where cohorts are defined on the basis of the year (spring) in which an individual should have been in third grade. The table shows that the 1998 cohort is the most recent one for which we are able to observe correlations between test scores and college-going. However, the 1998 cohort lacks third grade test score data. Earlier cohorts (1994 to 1997) have tests available in few grades, and later cohorts (1999 to 2008) lack college-going information altogether. In the present analysis, we use all available data, and allow the variance of the endowment (σ_{μ}^{2}) to vary across cohorts in our preferred specifications.

The CMS data present two additional challenges for our analysis. First, the EOG math exams experienced some changes in content and in testing process over the 1999 to 2009 period.[12] In our model, such changes should affect the correlation between test scores administered in separate testing regimes through a change in the test's reliability ratio. Similarly, there were some changes in district demographics over the period, arising both from Hispanic immigration and white return to the district after the end of court-ordered school desegregation in 2001. If test reliability depends on the underlying variance in achievement, these changes could also affect the reliability ratio. To account for all of these factors, we ultimately allow the reliability ratio (λ_{y}) to vary year-by-year (not just when the tests themselves change), in addition to varying across grades.

[12] In particular, the math test was changed in 2001, then again in 2006. Changes in the math test editions affected the weights given to different topics. The main change in the testing process over the period was the addition of more time with each edition; the current edition is untimed.

Second, our model assumes that year and grade move in lockstep, one-for-one. However, some children repeat grades, while others (albeit a much smaller fraction) skip ahead. We would like to use these children's scores, but we are uncomfortable assuming, for example, that a child repeating fifth grade and thus administered the fifth grade EOG math exam would occupy the same place in the distribution of sixth grade EOG math scores. To confront this complication, we estimated our model using the subset of students who never repeated (or skipped) a grade. This uses a consistent sample to calculate all correlations.

The first column of Table I presents demographic characteristics and college-going rates of the CMS students used to create these correlations.[13] CMS students are more likely to be minorities and have lower college attendance rates than we see in a random sample of U.S. twenty year olds in 2007, as shown in the third column based on the American Communities Survey (ACS).[14] This is not surprising, since CMS is an urban school district. The CMS students in our sample are also slightly more likely to be female than the population at large. This finding is expected, since we have restricted calculation of the correlations to individuals who have not repeated grades, and boys have higher grade repetition rates.

B. *Data from the National Longitudinal Survey of Youth*

The National Longitudinal Survey of Youth 1979 (NLSY79) is a widely-used, nationally representative longitudinal survey of roughly 12,000 individuals who were between the ages of 14 and 21 in 1979. In 1986, the survey incorporated data on the children born to female respondents, and the children have been surveyed every two years ever since, with most recent data from 2008. The resulting data set – the NLSY79 Child and Young Adult survey – includes biannual scores on

[13] The unit of observation is a correlation. For each correlation, we first calculate the demographic characteristics of the underlying sample. We then take the unweighted average of these means to arrive at the figures presented in the table. The figures are very similar if we weight by the number of observations used to create each correlation.

[14] Twenty-year-olds in 2007 would have been born in 1987 and should have entered third grade in 1995 and 1996. The sample therefore includes individuals from a cohort that is represented in both the CMS and NLSY79 Child-Mother data. The NLSY79 sample is described in the next section.

Peabody Individual Achievement Tests (PIAT) of math while a child is between the ages of 5 and 14, and information on whether he or she attended college after high school. For the latter, we rely on questions on educational attainment fielded in the young adult questionnaire, which began in 1994, in the survey years when an individual would have turned 20 or 21, depending on the cohort.

One benefit of these data over the CMS data is that test scores are available over a longer span. It is also the case that we can observe all possible correlations between test scores in different grades and between test scores and college going for a number of cohorts. We show available data by cohort in Appendix Table II; cohort now corresponds to birth year. We restrict our analysis sample to individuals born between 1982 and 1987, for which we have the complete span of test scores and the information on college-going after high school described above. Although individuals born in 1981 and prior satisfy these same criteria, we exclude them because they are born to increasingly younger mothers given the survey design. As is the case in our analysis of the CMS data, we take account of cohort differences in knowledge by allowing the endowment variance (σ_{μ}^{2}) to vary across cohorts.

The second column of Table I shows demographic characteristics and college-going rates of individuals in our sample from the NLSY79 Child and Young Adult survey.[15] Despite excluding the oldest children surveyed, our estimation sample – like that for CMS – is high minority and has low college attendance relative to the national average figures from the ACS. Our estimates may therefore not necessarily reflect what we would see had we access to nationally representative data. This is not necessarily a drawback of our analysis, however, as most educational interventions target more disadvantaged populations.

The NLSY79 has several other features that are useful for our analysis. Compared to the EOG math tests in CMS, test content and scoring were consistent over the period relevant for the

[15] The correlations are calculated in the same way they were for the CMS data.

cohorts in our sample (1986 to 2002), though there were some changes in test administration and the tested population over this span.[16] Again, we account for the effects of these changes by allowing test reliability to change over time in addition to across grades in some specifications. Further, children of the same age are administered the same test regardless of their grade of enrollment. There is therefore no need to restrict attention to non-repeaters.

The main drawbacks of the NLSY79 data are that the samples used to calculate any given correlation are quite small relative to those available for CMS,[17] and data are collected only every two years, not annually. This added noise to our estimates could potentially outweigh any increases in precision from having a wider grade span represented in the data.

V. Estimates of the Model Parameters

Table II presents equal-weighted minimum distance estimates of the model parameters based on the correlation data derived from CMS and the NLSY79, respectively. Panel A of the table pertains to the estimates based on CMS data, while Panel B corresponds to estimates based on the NLSY79.[18] To establish a benchmark, we begin with a parsimonious specification. In column (1), we force the innovation variance to be unchanging across grades, or $\beta_v = 1$, and assume that the reliability ratio for the test is unchanging across grades and years. The estimates of δ from this specification are above one in both datasets, though not significantly so in the CMS data. Allowing estimates of σ_μ^2 to vary across cohorts (column (2)) improves model fit (in the case of the CMS data) and reduces estimates of δ, though these estimates remain above one in both datasets. The specifications in the remaining columns of the table examine whether this finding continues to hold when we place fewer restrictions on the model estimated.

[16] Most notably, there was a shift to computer aided test administration in 1994. This apparently increased the proportion of children with valid scores by reducing interviewer error in test administration. There was also a move to testing in the English language only in 2002, and the minority oversample of the NLSY79 was also excluded from the 2000 wave (Center for Human Resource Research, 2009).

[17] On average around 350 observations are used to calculate each correlation – 15 times less than in the CMS data.

[18] Heteroskedasticity-robust standard errors are in parentheses.

Before turning to discussion of these next specifications, it is useful to note that estimates of the reliability ratios, λ_y and λ_w, are very precisely estimated and have magnitudes that line up with expectations. In the models estimated on the correlations for CMS, for instance, the value of λ_y =0.883 to 0.885 is consistent with reported reliabilities on the North Carolina EOG math test of 0.85 to 0.90. The estimate of λ_y =0.704 in the NLSY79 data is at the low end of the range of reliability estimates reported for the PIAT math test, but this may reflect noise that arises from other sources unique to the NLSY79, such as inconsistent child effort in a non-school testing environment.[19] In both samples, the estimates of λ_w are also much lower than estimates of λ_y, implying that accumulated knowledge contributes considerably less power in explaining the variation in college-going than the variation in test scores. Our estimates of λ_w in the NLSY79 are consistent with Neal and Johnson's (1996) findings that performance on the Armed Forces Qualifying Test as an adolescent explains roughly 15 percent of variation in earnings. Estimates of λ_w maybe be lower in the NLSY79 because self-reports of college-going are noisier than the administrative data on college-going available for CMS students. In any event, the finding of a relatively low reliability ratio for college-going reminds us that college-going is determined by much more than the notion of achievement that is central to student assessment.

Moving forward, we next (column (3)) allow the reliability ratio for the test to vary across grades. Recall that we are able to do this because test scores in a given grade are correlated against test scores in multiple other grades; our estimates attempt to pick up whether all correlations with

[19] In our model, λ_y is not directly comparable to the usual test-retest reliability because short-term shocks to knowledge (ω_{it}) are counted as noise for the purpose of estimating λ_y (see equation (7)). If we assume that the variance in the short-run shocks is roughly equal to the variance in the long-run shocks (consistent with about half of value added to knowledge fading by the next year) then our estimates of λ_y are consistent with test-retest reliability rates of just over .9 for the CMS data and around .75 for the NLSY79 data, which are still consistent with reported reliabilities.

test scores in, say, kindergarten are systematically low. Estimates of λ_y are now estimates of the reliability of the third grade test; the other coefficients presented are the predicted *differences* in test reliability between the grade specified and grade three. This specification improves model fit in both data sets, as reflected in the fall in the root mean squared error (RMSE) between columns (2) and (3). In Panels A and B alike, the estimates of test reliability are also roughly consistent with expectations: test reliability tends to improve across grades. The improvement in test reliability between kindergarten and third grade in the NLSY79 Child-Young Adult data is particularly striking: our estimates imply that kindergarten and first grade test scores are highly unreliable, with reliability ratios of 0.371 and 0.481, respectively. Increases in test reliability across grades will tend to make the statistical artifact less important; we show this empirically below. Allowing baseline (third grade) test reliability then to vary across years (column (4)) does not alter this pattern of findings and only moderately improves model fit (more so in the CMS data, where the test changed across years, than in the NLSY, where the same test was used in all years) .

Estimates of δ in the specifications with changing reliability are lower than those in columns (1) and (2) for both data sets, suggesting that changes in test reliability were previously loading onto this parameter. Allowing only for changes in test reliability across grades, in column (3), has the largest impact on estimates of δ . This is intuitive, since increases in test reliability across grades will strengthen the correlation between test scores in higher grades, creating the illusion of more knowledge accumulation as a child ages. For the specification in column (4), the estimates of δ from the two datasets are nearly identical, and imply decay in long-term knowledge of 6 to 7 percent per year. Nevertheless, estimates of δ remain statistically indistinguishable from one in both data sets.

To provide a sense of model fit, Figure III and Figure IV plot average values of the correlations in our data, as well as average *predicted* correlations based on our estimates from the specification in Table II, column (4) for the CMS and the NLSY79 data, respectively. Panel A of each figure is arranged like Figure I, where the grade of the earlier test is represented on the x-axis and the grade of the later test is represented by that value plus the number to the right of each locus. Panel B of each figure plots the correlation between test scores and college-going and is arranged like Figure II, with the grade of the test on the x-axis. In each figure and panel, the free-standing solid markers represent the actual correlations, and the hollow markers attached by dashed lines represent the predicted correlations. Note that the predicted correlations do not evolve smoothly as in Figures I and II, since the specification in column (4) allows for changes in the reliability ratio across grades and over time. Allowing for this heterogeneity generates predictions closer to the data, or improved model fit. In general, the model appears to fit the data quite well.

For completeness, the remaining column in each panel of Table II presents estimates from an even less restrictive specification. Here, we loosen the requirement of constancy in the variances of the innovations to long-run knowledge. We find no evidence to suggest that the variance of the innovations changes: the estimates of β_v are 0.947 and 0.985 for the CMS and NLSY79 samples, respectively, and in neither case can we reject the null that the variance is unchanging ($\beta_v = 1$). Estimates of δ , λ_w , and the grade-specific reliabilities are little changed from column (4). Moving forward, we therefore consider column (4) to present our preferred estimates. Despite considerable differences in the nature of the tests and the populations tested, this specification yields remarkably similar estimates of key parameters across the two data sets.

VI. Estimates of the Policy Parameters

Recall that the main objective of our analysis is to estimate the "statistical artifact"– how the scaling of test scores changes across grades due to differences in the ratio of the standard deviation in test scores in different grades. The statistical artifact is the product of two quantities: (1) the ratio of the standard deviation in long-run knowledge in different grades, or $\sqrt{\mathrm{var}\left(g_{it}^{L}\right)/\mathrm{var}\left(g_{it+k}^{L}\right)}$; and (2) the ratio of the (square root of the) inverse reliability ratios in different grades, or $\sqrt{\left(1/\lambda_{y_t}\right)/\left(1/\lambda_{y_{t+k}}\right)}$. We make our initial calculations under the assumption that test reliability is unchanging – or we focus initially on (1) – since variation in test reliability across grades will be idiosyncratic to the application. Because tests appear to become more reliable as children age, however, these initial estimates are likely an upper bound on the statistical artifact. We show this by calculating the statistical artifact using our estimates of test reliability.

To proceed, we need to choose the grade to which to compare. Because it allows us to describe how the variance in accumulated knowledge evolves over the entire elementary and secondary career, and because it provides a useful point of reference to the literature, we start by choosing kindergarten (t=0) as the comparison grade. Columns (1) and (4) of Table III present, for the CMS data and NLSY79 data, respectively, estimates of the statistical artifact under this assumption and the assumption of constant reliability, i.e., estimates of $\sqrt{\mathrm{var}\left(g_{i0}^{L}\right)/\mathrm{var}\left(g_{ik}^{L}\right)}$. In both datasets, the estimates are uniformly below one, and significantly so, suggesting that the variance in accumulated knowledge is indeed higher in subsequent grades than it is in kindergarten. The estimates are also monotonically decreasing in grade, implying that the variance in accumulated knowledge is increasing as children age. Furthermore, knowledge accumulates most rapidly in the earliest grades: in both datasets, the variance in accumulated knowledge rises by slightly more between kindergarten and fourth grade than it does between grades four and twelve. To see this, note that the variance in knowledge in each grade, relative to kindergarten, is the squared inverse of the estimates in columns (1) and (4). Using the CMS estimates, the variance of accumulated knowledge is 19 percent larger in fourth grade and 37 percent larger in twelfth grade than it was in kindergarten. Using the NLSY79 estimates, these figures are larger, at 29 percent and 56 percent, respectively.

In the next columns ((2) and (5)), we use third grade instead of kindergarten as the comparison grade. We choose grade three because it is observed in both datasets – thereby allowing us to recalculate the statistical artifact accounting for changing reliability – and it is the earliest grade observed for CMS. Consistent with the observations from columns (1) and (4), the variance in knowledge does not rise dramatically from grade three forward. Accounting for changing reliability (columns (3) and (6)), the variance in test scores increases by even less. In the NLSY79, where we can estimate test reliability in earlier grades, increasing reliability of tests between kindergarten and third grade is large enough to offset the growing variance of knowledge, resulting in an overall decline in the variance of test scores until third grade. These results suggest that relative comparisons of intervention effects at very young ages depend importantly on the relative reliability of the early tests.

We originally motivated our analysis with two empirical regularities: interventions in later grades tend to have smaller effects than the same interventions in earlier grades, and the test score impacts of early educational interventions almost universally "fade out" over time. The remainder of this section focuses on the implications that our results have for interpreting these two empirical regularities.

A. Implications for Contemporaneous Effects of Interventions in Different Grades

The statistical artifact conveys how much smaller the test score impacts of an intervention in a later grade will be when compared to the test score impacts of an intervention in kindergarten that has the same impact on knowledge (equation (11)). That is, suppose that an intervention in kindergarten raises kindergarten test scores by 0.2 standard deviations. The estimates in column (1) of Table III imply that an intervention with the same impact on knowledge will raise test scores by only 0.175 standard deviations in eighth grade, according to the CMS (0.2*0.877), and by only 0.166 standard deviations, according to the NLSY79 (0.2*0.828). Thus, the estimated effects on test scores would shrink by 12 to 17 percent despite no change in the effects on knowledge. Using the estimates in columns (2) and (5), the estimated effects on test scores of two identical interventions would shrink by 6 to 8 percent between grade three and grade eight. Allowing for changing test reliability reduces the expected reduction in effects to less than 4 percent, based on the CMS estimates (column (3)), and raises it to about 10 percent, based on the NLSY79 estimates (column (6)).

These estimates are too small to explain fully the differences in the effects of the same intervention in different grades. For example, Dee and Jacob (2011) estimate that the No Child Left Behind Act (NCLB) had an effect on math test scores in eighth grade that was smaller than that effect in fourth grade by 58 percent or more. Likewise, in a study of charter schools in the Harlem Children's Zone, Dobbie and Fryer (2011) generally find that the achievement effects of a year of charter attendance in elementary school exceeds those in middle school, and by a factor much larger that implied by Table III. Finally, Kane, Rockoff, and Staiger (2008) estimate that differences across teachers in their impact on student test scores were about 40 percent lower in middle school than in elementary grades.[20]

While the standard errors on all of these estimates might admit changes in test scores as small as those implied by Table III, the more plausible explanation is that the same intervention later in the school career has less of an impact on knowledge. Still, the statistical artifact is economically meaningful. For example, taking grades three and eight to represent elementary and middle school,

[20] A similar set of findings also holds in studies of the net impacts of additional year of completed education on test scores. For example, exploiting the sharp difference in years of school completed among children with birthdays near school entry cutoff dates, Anderson et al. (2011) estimate that exposure to first grade raises math achievement by 0.776 standard deviations. Using a similar identification strategy, Cascio and Lewis (2006) find that an additional year of education raises the achievement of minority teenagers by 0.3 to 0.4 standard deviations.

respectively, our estimates from CMS imply that at least 10 percent of the decline in teacher effects is a statistical artifact. If the artifact is ignored, policymakers seeking to compare teachers in different grades on the basis of their value-added will systematically reward equally-effective teachers in earlier grades.

B. *Implications for Interpretation of Fadeout*

An alternative way to interpret the statistical artifact is that it answers the counterfactual question: how much fadeout would be observed by grade *k* if the math knowledge gained from an intervention in kindergarten persisted in full; that is, what would θ_k^0 / θ_0^0 be if $\delta = 1$ and $\tau = 0$, i.e., if all knowledge gains were long term and did not fade? Again using eighth grade (*k*=8) as an example, the answer to this question is 12 to 17 percent, depending on the dataset. While concise, this explanation is a bit misleading, since estimates of $\sqrt{\mathrm{var}\left(g_{i0}^L\right) / \mathrm{var}\left(g_{ik}^L\right)}$ depend on estimates of δ, and τ may be non-zero. So more precisely, the statistical artifact conveys how much more fadeout is observed than would be predicted by true knowledge decay alone. By eighth grade, the answer to this question would be 12 to 17 percent.

In practice, fadeout of only 12 to 17 percent by eight years after an intervention is low. For example, Krueger and Whitmore (2001) estimate the effects on attending a small class in Project STAR on test scores through eighth grade.[21] While several issues complicate comparison of their estimates to ours,[22] one cannot deny that they find greater fade-out – on the order of 70 to 80 percent – by eighth grade. Moreover, most of the fadeout in effects on test scores in Project STAR occurs immediately after the class size intervention ends. Similarly, estimates of fade-out in teacher

[21] Project STAR (Student Teacher Achievement Ratio), also known as the Tennessee STAR experiment, randomly assigned children entering kindergarten in 1985 to small and regular sizes in roughly 80 Tennessee schools. Children entering these schools in later grades were also randomly assigned to class types, and participants were expected to stay in their initially assigned type of class through third grade, after which the experiment ended.

[22] We normalize tests in standard deviation units and look at math, while they use percentile ranks and average across performance on math and reading. Further, the treatment was not confined just to kindergarten, but spanned up to four years for some students.

effects, which generally rely on data from third grade and later, reach 50 percent or higher only one year out (Kane and Staiger, 2008; Jacob, Lefgren, and Sims, 2008; Rothstein, 2010; Chetty, Friedman, and Rockoff, 2011). As was shown in Table III, the statistical artifact is even less important for understanding fadeout from interventions in third grade forward (columns (2) and (5)), particularly once test reliability is taken into account (columns (3) and (6)).

Yet, our estimates can be reconciled with this pattern of findings from the literature. While our model predicts a small degree of fadeout from the statistical artifact, it can generate *overall* fadeout of roughly this magnitude. To see this, recall that our model allowed for two sources of true knowledge decay, embodied in the parameters δ and τ . δ captures the persistence of innovations to long-run knowledge, v_{it}. Our preferred specification from Table II, column (4) estimates δ at 0.935 or 0.936, implying that innovations to long-run knowledge fade by only 6 to 7 percent per year. τ , in turn, is the share of the intervention's contemporaneous effect that does not carry over to later grades (equation (10)), which depends on an intervention's effect on ω_{it} – the transitory innovation to knowledge – relative to its effect on v_{it}.

Unfortunately, we cannot estimate τ directly from our data, and it may vary depending on the intervention. As discussed earlier, the prior literature suggests that half or more of the effects of teachers and other interventions fade after the first year, suggesting that a value of $\tau = 0.5$ seems reasonable. Alternatively, we can use the reported reliability ratios for the tests in CMS (0.9 or lower) – which account only for testing noise – to put an upper bound on the variance of the transitory innovations to knowledge (which our model does not separately identify from the variance of testing noise, var(ε_{it}), so it is subsumed in our estimates of the reliability ratio, λ_{y_t}). Using this upper bound on var(ω_{it}), and taking advantage of the fact that our specification can be rewritten as an ARMA(1,1) in which the MA parameter corresponds to τ (see footnote 7), we can get a sense of a reasonable value for τ. These back-of-the-envelope calculations suggest τ on the order of 0.3 to 0.4.[23]

Assuming a mid-range value of $\tau = 0.4$, Figure V plots estimates of total fadeout against *k* from an intervention in kindergarten for the CMS data (upper row) and NLSY79 data (lower row), respectively. The underlying estimates of model parameters correspond to the preferred specification, presented in column (4) of Table I. The solid lines connect the estimates themselves, and the dotted lines represent their 95 percent confidence intervals.[24] The first column of each row plots the total fadeout from all sources – both true knowledge decay, embodied in the parameters δ and τ, and the statistical artifact. The overall pattern of fadeout in these figures is consistent with the general pattern observed in the literature, with an initial sharp decline of about half, followed by a more gradual decline toward one quarter or one third of the original effect. For the purposes of comparison, the second column of each row presents estimates of the statistical artifact (from columns (1) and (4) of Table III), and the third column presents the percent of total fadeout that is a statistical artifact. The statistical artifact accounts for little of the short-run fadeout in test scores, but accounts for approximately 20 percent of total fadeout after four to five years.

VII. Conclusion

Interventions in education are frequently found to have effects on test scores that fade over time: those that occur earlier tend to have larger test score impacts than those that occur later, and the test score impacts of earlier interventions do not appear to last. This paper has examined

[23] Based on the estimates from our preferred specification for CMS in column 4 of Table II, values of var(ω_{it}) (relative to the variance of the AR(1) innovations in long-term knowledge, which are normalized to 1) of .6 to 1 (depending on grade) are consistent with conventional test reliability rates of .9. This implies that short-term plus long-term knowledge in our model (excluding the fixed component) is the sum of an AR(1) with AR parameter .936 and innovation variance of 1, and a white noise term with variance .6 to 1. Using results from Granger and Morris (1976), this model is equivalent to an ARMA(1,1) with AR parameter .936 and MA parameter between -.29 and -.38. The MA parameter explicitly represents the proportion of this periods innovation that disappears immediately, as opposed to the geometric decay after the first period driven by the AR parameter.

[24] Standard errors were calculated using the delta method. Given that the underlying specification estimated different values of σ_{μ}^{2} for each cohort, we used the weighted average of the cohort-specific variances, where the weight is the frequency with which the cohort contributes to the estimation sample.

whether this pattern of findings can be explained by the common practice of rescaling test scores in distributional terms in an environment where the variance in knowledge is rising as children progress through school. Intuitively, if this variance grows, the same increase in knowledge will change a child's position in the achievement distribution by less in later grades: the effect on test scores falls, but the effect on knowledge does not – a "statistical artifact." We estimate the statistical artifact by fitting the predictions of a standard model of education production to correlations in test scores across grades and with college-going using both administrative and survey data.

Our findings suggest that the variance in knowledge does indeed increase in a statistically and economically significant way from the start of elementary school to the end of secondary school – by between 37 and 56 percent in our relatively unrestricted preferred specification, depending on the data set. Correspondingly, our estimates of the statistical artifact imply, for example, that that the test score impacts of attending a small class in kindergarten would be 12 to 17 percent lower by eighth grade even if the effect of a small kindergarten class on knowledge persisted in full. While the *overall* fadeout of an intervention in kindergarten observed in practice – and predicted by our model – is much greater, the statistical artifact is still substantial, representing approximately 20 percent of overall fadeout over the longer term. To take another example, our estimates imply that a teacher assigned to an eighth grade class would increase her students' test scores by 6 to 10 percent less than if she were assigned to teach third grade. Again, while the statistical artifact does not account for *all* observed decline in contemporaneous teacher effects, it is still substantial, particularly from the perspective of policymakers seeking to evaluate teachers from different grades on the basis of their value-added.

Our findings also have wider implications. This is perhaps most notably the case for research on racial, ethnic, and socioeconomic achievement gaps, which are commonly measured in distributional terms. Our findings suggest a standard deviation difference in test scores translates into a somewhat larger difference in knowledge in higher grades, with the implication that gaps in knowledge expand more rapidly as children progress through school than any expansion in the normalized test score gap (see, for example, Fryer (2011)) would imply.

References

Almond, Douglas and Janet Currie. 2011. "Human Capital Development Before Age Five." *Handbook of Labor Economics*, Volume 4, Part B, pp. 1315-1486.

Anderson, Patricia, Kristin Butcher, Elizabeth Cascio, and Diane Whitmore Schanzenbach. 2011. "Is Being in School Better? The Impact of School on Children's BMI When Starting Age is Endogenous." *Journal of Health Economics* 30(5): 977-986.

Ballou, Dale. 2009. "Test Scaling and Value-Added Measurement." *Education Finance and Policy* 4(4): 351-383.

Barlevy, Gadi and Derek Neal. 2011. "Pay for Percentile." *American Economic Review*, forthcoming.

Bond, Timothy N. and Kevin Lang. 2011. "The Evolution of the Black-White Test Score Gap in Grades K-3: The Fragility of Results." Mimeo, Boston University.

Cascio, Elizabeth and Ethan Lewis. 2006. "Schooling and the Armed Forces Qualifying Test: Evidence from School-Entry Laws." *The Journal of Human Resources* 41(2): 294-318.

Cawley, John, James Heckman and Edward Vytlacil, 1999. "On Policies to Reward the Value Added of Educators." *The Review of Economics and Statistics* 81(4):720-727.

Center for Human Resource Research. 2009. *NLSY79 Child & Young Adult Data User Guide: A Guide to the 1986-2006 Child Data 1994-2006 Young Adult Data.* Columbus, Ohio: The Ohio State University.

Chetty, Raj, John N. Friedman, Nathaniel Hilger, Emmanuel Saez, Diane Whitmore Schanzenbach, and Danny Yagan. 2011. "How Does Your Kindergarten Classroom Affect Your Earnings? Evidence from Project STAR." *Quarterly Journal of Economics* 126(4): 1593-1660.

Chetty, Raj, John Friedman, and Jonah Rockoff. 2011. "The Long-Term Impacts of Teachers: Teacher Value-Added and Student Outcomes in Adulthood." NBER Working Paper 17699. Cambridge MA: National Bureau of Economic Research.

Cunha, Flavio and James Heckman. 2008. "Formulating, Identifying and Estimating the Technology of Cognitive and Noncognitive Skill Formation." *Journal of Human Resources* 43(4):739-780.

Dee, Thomas S. and Brian Jacob. 2011. "The Impact of No Child Left Behind on Student Achievement." *Journal of Policy Analysis and Management* 30(3): 418-446.

Dobbie, Will and Roland Fryer. 2011. "Are High-Quality Schools Enough to Increase Achievement Among the Poor? Evidence from the Harlem Children's Zone." *American Economic Journal: Applied Economics* 3(3): 158-187.

Fryer, Roland. 2011. "Racial Inequality in the 21st Century: The Declining Significance of Discrimination." *Handbook of Labor Economics*, Volume 4, Part B, pp. 855-971.

Granger, C.W.J. and M.J. Morris. 1976. "Time Series Modelling and Interpretation." *Journal of the Royal Statistical Society. Series A (General)* 139(2): 246-257.

Hambleton, R. K., Swaminathan, H. and Rogers, H. J. 1991. *Fundamentals of Item Response Theory.* Newbury Park, CA: Sage Press.

Jacob, Brian, Lars Lefgren, and David Sims. 2008. "The Persistence of Teacher-Induced Learning Gains." *Journal of Human Resources*, forthcoming.

Kane, Thomas J. and Douglas O. Staiger. 2008. "Estimating Teacher Impacts on Student Achievement: An Experimental Evaluation." Mimeo, Harvard University and Dartmouth College.

Kane, Thomas J., Jonah E. Rockoff, and Douglas O. Staiger. 2008. "What Does Certification Tell Us About Teacher Effectiveness? Evidence from New York City." *Economics of Education Review* 27(6): 615-631.

Krueger, Alan and Diane Whitmore. 2001. "The Effect of Attending a Small Class in the Early Grades on College-Test Taking and Middle School Test Results: Evidence from Project STAR." *The Economic Journal* 111: 1-28.

Lang, Kevin. 2010. "Measurement Matters: Perspectives on Education Policy from an Economist and School Board Member." *Journal of Economic Perspectives* 24(3): 167-182

Lockwood, J. R., Daniel F. McCaffrey, Louis T. Mariano, and Claude Setodji. 2007. "Bayesian Methods for Scalable Multivariate Value-Added Assessment." *Journal of Educational and Behavioral Statistics* (32)*:*125 - 150.

McCaffrey, Daniel F., J. R. Lockwood, Daniel Koretz, Thomas A. Louis, and Laura Hamilton. 2004. "Models for Value-Added Modeling of Teacher Effects." *Journal of Educational and Behavioral Statistics* 29(1):67-101.

Neal, Derek A. and William R. Johnson. 1996. "The Role of Premarket Factors in Black-White Wage Differences. *The Journal of Political Economy* 104(5): 869-895.

Reardon, Sean F. 2007. "Thirteen Ways of Looking at the Black-White Test Score Gap." Mimeo, Stanford University.

Reardon, Sean F., & Robinson, J.P. 2007. "Patterns and trends in racial/ethnic and socioeconomic achievement gaps." In Helen A. Ladd & Edward B. Fiske (Eds.), *Handbook of Research in Education Finance and Policy*. Lawrence Erlbaum.

Rothstein, Jesse. 2010. "Teacher Quality in Educational Production: Tracking, Decay, and Student Achievement." *Quarterly Journal of Economics* 125(1): 175-214.

Selzer, Michael H., Ken A. Frank, and Anthony S. Bryk. 1994. "The metric matters: the sensitivity of conclusions about growth in student achievement to choice of metric." *Educational Evaluation and Policy Analysis* 16:41-49.

Table I. Demographic Characteristics of Underlying Samples

	Analysis Data:		*For Comparison:*
	Charlotte-Mecklenburg Schools Data	NLSY79 Child and Young Adult Data	American Community Survey, 2007
	(1)	(2)	(3)
Female	0.531	0.510	0.477
Non-black, non-Hispanic	0.533	0.461	0.668
Black	0.377	0.318	0.142
Hispanic	0.072	0.220	0.170
Multi-racial	0.018	n/a	0.021
College Attendance	0.484	0.508	0.600

Notes: In columns (1) and (2), the unit of observation is the correlation. For each correlation, we calculate the demographic characteristics of the underlying sample. The table presents unweighted averages of these demographic characteristics; figures are very similar if we weight by the number of observations used to generate the correlation. Means in column (3) are based on the sample of 36,252 20 year olds in the 2007 ACS and are calculated using population weights provided in the data.

Table II. Estimates of Model Parameters

	A. Charlotte-Mecklenburg Schools Data					B. NLSY79 Child and Young Adult Data				
	(1)	(2)	(3)	(4)	(5)	(1)	(2)	(3)	(4)	(5)
σ_μ^2 (natural log)	2.424					-20.12				
	(0.152)					(0)				
δ	1.060	1.019	0.916	0.936	0.922	1.164	1.051	0.908	0.935	0.933
	(0.0353)	(0.0278)	(0.0432)	(0.0330)	(0.0395)	(0.0242)	(0.0166)	(0.0584)	(0.0518)	(0.0573)
λ_w	0.289	0.272	0.279	0.283	0.276	0.167	0.207	0.208	0.206	0.203
	(0.00578)	(0.00515)	(0.00548)	(0.00581)	(0.00941)	(0.0119)	(0.0165)	(0.0172)	(0.0162)	(0.0228)
λ_y	0.883	0.885	0.865			0.704	0.743	0.718		
	(0.00519)	(0.00434)	(0.00436)			(0.0135)	(0.0153)	(0.0352)		
+ kindergarten $\Delta \lambda_y$								-0.347	-0.341	-0.338
								(0.0492)	(0.0497)	(0.0514)
+ 1st grade $\Delta \lambda_y$								-0.237	-0.227	-0.224
								(0.0359)	(0.0381)	(0.0403)
+ 2nd grade $\Delta \lambda_y$								-0.0717	-0.0725	-0.0704
								(0.0470)	(0.0401)	(0.0409)
+ 4th grade $\Delta \lambda_y$			0.0103	0.0150	0.0112			0.0263	0.0231	0.0233
			(0.00545)	(0.00508)	(0.00735)			(0.0376)	(0.0311)	(0.0317)
+ 5th grade $\Delta \lambda_y$			0.0184	0.0242	0.0194			0.0514	0.0405	0.0391
			(0.00656)	(0.00597)	(0.00909)			(0.0293)	(0.0304)	(0.0310)
+ 6th grade $\Delta \lambda_y$			0.0345	0.0412	0.0359			0.0186	0.00496	0.00418
			(0.00639)	(0.00579)	(0.00939)			(0.0367)	(0.0378)	(0.0394)
+ 7th grade $\Delta \lambda_y$			0.0433	0.0518	0.0456			0.00462	-0.00907	-0.0113
			(0.00542)	(0.00528)	(0.0102)			(0.0381)	(0.0349)	(0.0355)
+ 8th grade $\Delta \lambda_y$			0.0304	0.0416	0.0322			0.00429	-0.0264	-0.0292
			(0.00674)	(0.00648)	(0.0129)			(0.0416)	(0.0527)	(0.0568)
+ 9th grade $\Delta \lambda_y$								0.0931	0.0578	0.0521
								(0.0392)	(0.0442)	(0.0550)
β_v					0.947					0.985
					(0.0586)					(0.0974)
Root MSE	0.0150	0.0120	0.00955	0.00793	0.00795	0.0485	0.0508	0.0361	0.0341	0.0343
N (correlations)	145	145	145	145	145	90	90	90	90	90
Model:										
σ_μ^2 varies by cohort		X	X	X	X		X	X	X	X
λ_y varies by year				X	X				X	X

Notes: The models are estimated using equal weighted minimum distance. See text for description of the model and the data. Robust standard errors are in parentheses.

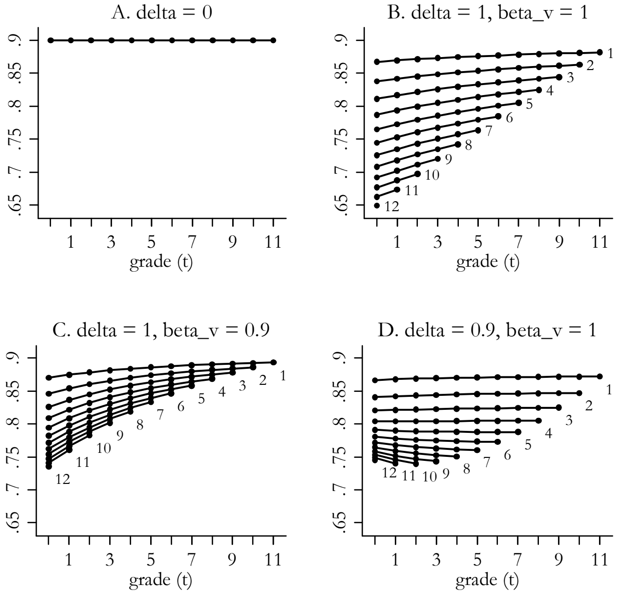
Table III. Estimates of Policy Parameters:

$$\sqrt{\frac{\left(1/\lambda_{y_t}\right)\text{var}\left(g_{it}^{L}\right)}{\left(1/\lambda_{y_{t+k}}\right)\text{var}\left(g_{it+k}^{L}\right)}}$$

		Charlotte-Mecklenburg Schools Data			NLSY79 Child and Young Adult Data		
	Assuming:	Constant λ_y		Changing λ_y	Constant λ_y		Changing λ_y
	t =	K (0)	3rd grade (3)	3rd grade (3)	K (0)	3rd grade (3)	3rd grade (3)
$t+k$ =		(1)	(2)	(3)	(4)	(5)	(6)
0	K	**1.000**			**1.000**	1.109	0.807
						(0.027)	(0.059)
1	1st grade	0.973			0.959	1.064	0.881
		(0.003)			(0.011)	(0.015)	(0.038)
2	2nd grade	0.950			0.927	1.028	0.975
		(0.007)			(0.017)	(0.007)	(0.033)
3	3rd grade	0.932	**1.000**	**1.000**	0.902	**1.000**	**1.000**
		(0.010)			(0.022)		
4	4th grade	0.917	0.984	0.992	0.881	0.977	0.993
		(0.014)	(0.004)	(0.006)	(0.026)	(0.007)	(0.021)
5	5th grade	0.904	0.970	0.984	0.864	0.958	0.985
		(0.017)	(0.009)	(0.011)	(0.030)	(0.014)	(0.025)
6	6th grade	0.894	0.959	0.981	0.850	0.943	0.946
		(0.021)	(0.013)	(0.015)	(0.034)	(0.021)	(0.030)
7	7th grade	0.885	0.949	0.977	0.838	0.930	0.924
		(0.025)	(0.017)	(0.019)	(0.039)	(0.027)	(0.032)
8	8th grade	0.877	0.941	0.963	0.828	0.919	0.902
		(0.028)	(0.021)	(0.023)	(0.043)	(0.034)	(0.047)
9	9th grade	0.870			0.820	0.910	0.945
		(0.031)			(0.048)	(0.040)	(0.057)
10	10th grade	0.865			0.813		
		(0.035)			(0.053)		
11	11th grade	0.860			0.807		
		(0.038)			(0.057)		
12	12th grade	0.856			0.802		
		(0.040)			(0.061)		

Notes: Standard errors (calculated using the delta method) in parentheses. The underlying specification is that presented in column (4) of Table II. This specification allows σ^2_μ (the endowment variance) to vary across cohorts and λ_y (the reliability ratio for third grade test scores) to vary across years. To calculate var(g) in this table, we therefore must assume values for σ^2_μ (in all columns) for λ_y (columns (3) and (6)). For σ^2_μ, we use the weighted average of the cohort-specific variance estimates, where the weight is the relative frequency with which the cohort contributes to the sample. For λ_y, we use the weighted average of the year-specific third-grade reliability ratio estimates, where the weight is the relative frequency with which test scores from the year appear in the correlations.

Figure I. Simulations of corr(y(t),y(t+k))
Under Alternative Assumptions on Model Parameters


A. delta = 0

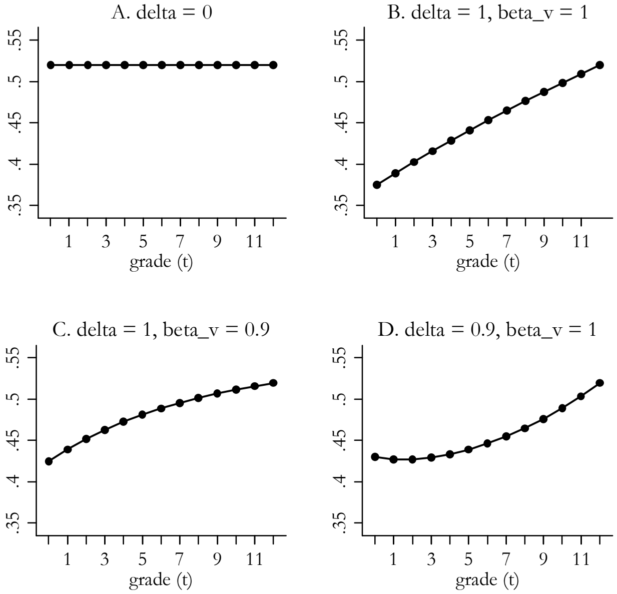
B. delta = 1, beta_v = 1

C. delta = 1, beta_v = 0.9

D. delta = 0.9, beta_v = 1

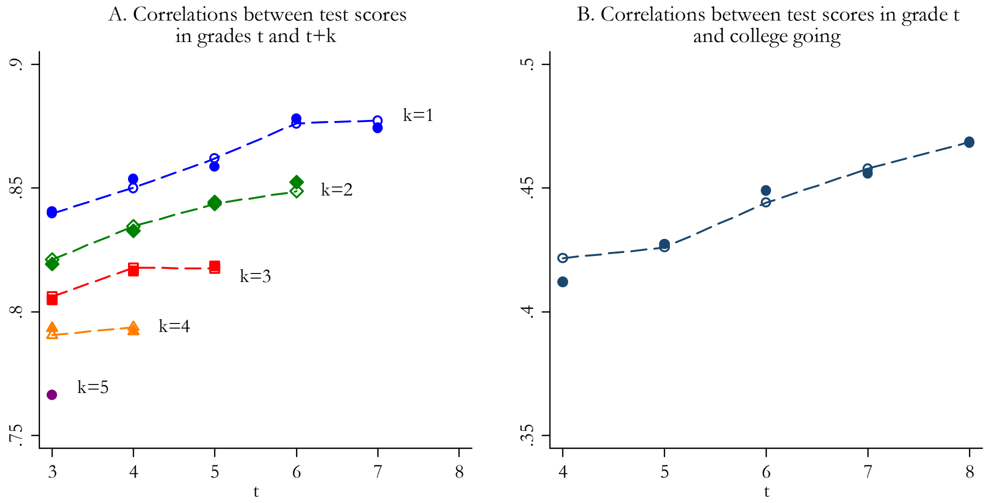
Note: The numbers plotted represent k. Setting lambda=.9, lambda_w=0.3, and sig2_mu=12.

Figure II. Simulations of corr(y(t),w)
Under Alternative Assumptions on Model Parameters



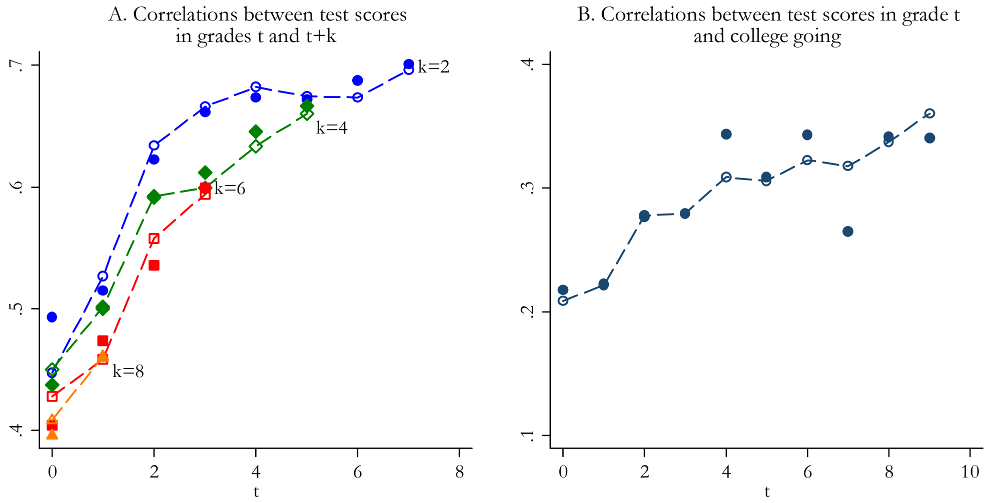
Note: Setting lambda=.9, lambda_w=0.3, and sig2_mu=12.

Figure III. Average Correlations and Average Predicted Correlations from Preferred Specification: the Charlotte-Mecklenburg Schools Data



Notes: The free-standing solid markers represent the (unweighted) average of the actual correlations, and the hollow markers attached by dashed lines represent the (unweighted) average of the predicted correlations. Predictions are from the model presented in column (4) of Table II.

Figure IV. Average Correlations and Average Predicted Correlations
from Preferred Specification: the NLSY79 Child and Young Adult Data



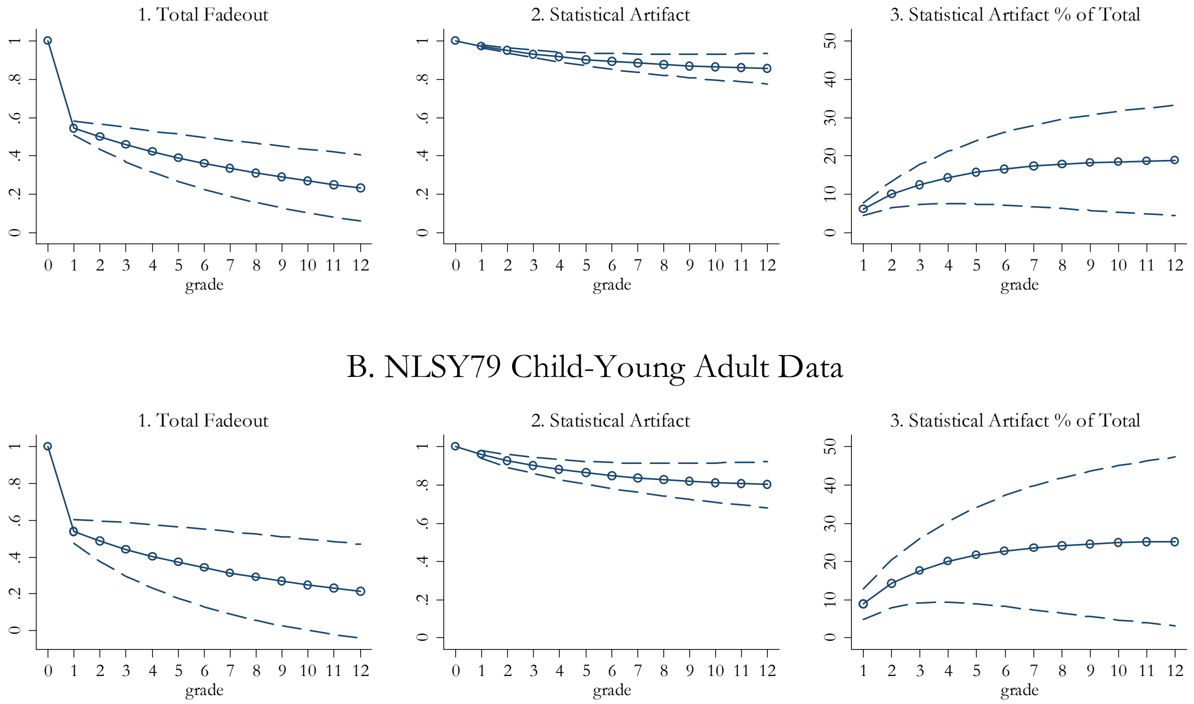
A. Correlations between test scores in grades t and t+k

B. Correlations between test scores in grade t and college going

Notes: See notes to Figure III.

Figure V. Comparison of Statistical Artifact to Total Fadeout: Intervention in Kindergarten

A. Charlotte-Mecklenburg Schools Data



B. NLSY79 Child-Young Adult Data

Notes: Graph 1 of each panel plots $\delta^k \cdot (1-\tau) \cdot \sqrt{(1/\lambda_{y_t})\mathrm{var}(g_{it}^L)/(1/\lambda_{y_{t+k}})\mathrm{var}(g_{it+k}^L)}$ under the assumption that $\tau = 0.4$, $\lambda_{y_t} = \lambda_{y_{t+k}}$, and t=0 (kindergarten). Graph 2 of each panel plots $\sqrt{(1/\lambda_{y_t})\mathrm{var}(g_{it}^L)/(1/\lambda_{y_{t+k}})\mathrm{var}(g_{it+k}^L)}$ under the assumption that $\lambda_{y_t} = \lambda_{y_{t+k}}$ and t=0. The underlying specification is that in Table II, column (4). This specification allows σ^2_μ (the endowment variance) to vary across cohorts. To calculate var(g) in this table, we set σ^2_μ as the weighted average of the cohort-specific variance estimates from this model, where the weight is the frequency with which the cohort contributes to the sample. The dashed lines represent 95 percent confidence intervals. Standard errors were calculated using the delta method.

Appendix Table I. Data Availability by Cohort: Administrative Data from Charlotte-Mecklenburg Schools

Year (Spring) in Grade:						
3	4	5	6	7	8	College 1
1994	1995	1996	1997	1998	1999	2004
1995	1996	1997	1998	1999	2000	2005
1996	1997	1998	1999	2000	2001	2006
1997	1998	1999	2000	2001	2002	2007
1998	1999	2000	2001	2002	2003	2008
1999	2000	2001	2002	2003	2004	2009
2000	2001	2002	2003	2004	2005	2010
2001	2002	2003	2004	2005	2006	2011
2002	2003	2004	2005	2006	2007	2012
2003	2004	2005	2006	2007	2008	2013
2004	2005	2006	2007	2008	2009	2014
2005	2006	2007	2008	2009	2010	2015
2006	2007	2008	2009	2010	2011	2016
2007	2008	2009	2010	2011	2012	2017
2008	2009	2010	2011	2012	2013	2018

Notes: The shaded cells correspond to years in which data are available for CMS. The test score data thus span 1999 2009, and college-going in the first year after high school is available through fall 2008. The present analysis uses data on all cohorts.

Appendix Table II. Data Availability by Cohort: National Longitudinal Survey of Youth 1979 Child-Young Adult Survey

	Year Turn Age:									
Year Born	5	6	7	8	9	10	11	12	13	14	20 or 21
1972	1977	1978	1979	1980	1981	1982	1983	1984	1985	1986	1992
1973	1978	1979	1980	1981	1982	1983	1984	1985	1986	1987	1994
1974	1979	1980	1981	1982	1983	1984	1985	1986	1987	1988	1994
1975	1980	1981	1982	1983	1984	1985	1986	1987	1988	1989	1996
1976	1981	1982	1983	1984	1985	1986	1987	1988	1989	1990	1996
1977	1982	1983	1984	1985	1986	1987	1988	1989	1990	1991	1998
1978	1983	1984	1985	1986	1987	1988	1989	1990	1991	1992	1998
1979	1984	1985	1986	1987	1988	1989	1990	1991	1992	1993	2000
1980	1985	1986	1987	1988	1989	1990	1991	1992	1993	1994	2000
1981	1986	1987	1988	1989	1990	1991	1992	1993	1994	1995	2002
1982	1987	1988	1989	1990	1991	1992	1993	1994	1995	1996	2002
1983	1988	1989	1990	1991	1992	1993	1994	1995	1996	1997	2004
1984	1989	1990	1991	1992	1993	1994	1995	1996	1997	1998	2004
1985	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2006
1986	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2006
1987	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2008
1988	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2008
1989	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2010
1990	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2010
1991	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2012
1992	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2012
1993	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2014
1994	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2014
1995	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2016
1996	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2016
1997	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2018
1998	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2018
1999	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2020
2000	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2020
2001	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2022

Notes: The shaded cells correspond to years in which either PIAT assessments were administered in the child survey or college-going information is reported in young adult survey, The child survey began in 1986 and the young adult survey began in 1994; both are administered biennially. The present analysis uses data on the 1982 to 1987 birth cohorts.